Empirical analysis of web-based user-object bipartite networks

Ming-Sheng Shang[1], Linyuan Lü[2], Yi-Cheng Zhang[1,2] and Tao Zhou[2,3(a)]

[1] *Web Science Center, School of Computer Science and Engineering, University of Electronic Science and Technology, 610054 Chengdu, P. R. China*
[2] *Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland*
[3] *Department of Modern Physics, University of Science and Technology of China, Hefei 230026, P. R. China*

PACS 89.75.Hc – Networks and genealogical trees
PACS 89.75.-k – Complex systems
PACS 89.20.Ff – Computer science and technology

Abstract. - Understanding the structure and evolution of web-based user-object networks is a significant task since they play a crucial role in e-commerce nowadays. This Letter reports the empirical analysis on two large-scale web sites, *audioscrobbler.com* and *del.icio.us*, where users are connected with music groups and bookmarks, respectively. The degree distributions and degree-degree correlations for both users and objects are reported. We propose a new index, named *collaborative clustering coefficient*, to quantify the clustering behavior based on the collaborative selection. Accordingly, the clustering properties and clustering-degree correlations are investigated. We report some novel phenomena well characterizing the selection mechanism of web users and outline the relevance of these phenomena to the information recommendation problem.

Introduction. – The last decade has witnessed tremendous activities devoted to the understanding of complex networks [1–5]. A particular class of networks is the *bipartite networks*, whose nodes are divided into two sets X and Y, and only the connection between two nodes in different sets is allowed. Many systems are naturally modeled as bipartite networks [6]: the human sexual network [7] consists of men and women, the metabolic network [8] consists of chemical substances and chemical reactions, the collaboration network [9] consists of acts and actors, the Internet telephone network consists of personal computers and phone numbers [10], etc. In addition to the empirical analysis on the above-mentioned bipartite networks, great effort has been made in how to characterize bipartite networks [11–13], how to project bipartite networks into monopartite networks [14–16] and how to model bipartite networks [17–20].

An important class of bipartite networks is the *web-based user-object networks*, which play the central role in e-commerce for many online selling sites and online services sites [21]. This class of networks has two specific evolving mechanisms different from the well-understood act-actor bipartite networks and human sexual networks. Firstly, connections between existent users and objects are generated moment by moment while this does not happen in act-actor networks (e.g., one can not add authors to a scientific paper after its publication). Secondly, users are active (to select) while objects are passive (to be selected). This is different from the human sexual networks where in principle both men and women are active. In

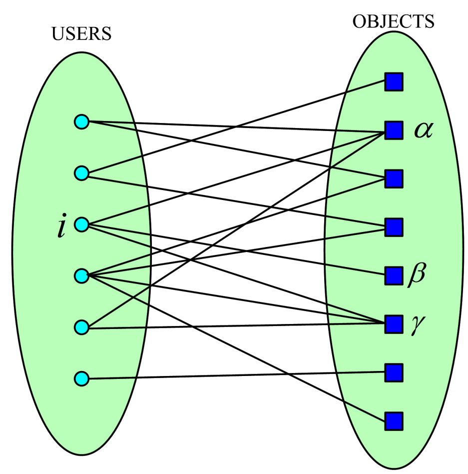

Fig. 1: (Color online) Illustration of a small user-object bipartite network.

(a) Corresponding author: zhutou@ustc.edu

Table 1: The basic properties of the two data sets. N, M and E denote the numbers of users, objects and edges, respectively. $\langle k\rangle$ and $\langle d\rangle$ are the average user degree and average object degree. C_u and C_o are the collaborative clustering coefficients for users and objects, and for comparison, $\bar{s_o}$ and $\bar{s_u}$ are the average similarities over all object pairs and over all user pairs, respectively. The user selection is considered to be highly clustered since $C_u \gg \bar{s_o}$.

Data	N	M	E	$\langle k\rangle$	$\langle d\rangle$	C_u	$\bar{s_o}$	C_o	$\bar{s_u}$
Audioscrobbler.com	35916	617900	5028580	140.01	8.14	0.0267	9.96×10^{-5}	0.0198	4.82×10^{-3}
Del.icio.us	10000	232658	1233995	123.40	5.30	0.0338	4.64×10^{-4}	0.0055	8.10×10^{-4}

a word, the user-object networks are driven by selection of users while the human sexual networks are driven by matches. Bianconi *et al.* [22] investigated the effects of the selection mechanisms of users on the network evolution. Lambiotte and Ausloos [23, 24] analyzed the web-based bipartite network consisted of listeners and music groups, especially, they developed a percolation-based method to uncover the social communities and music genres. Zhou *et al.* [15] proposed a method to better measure the user similarity in general user-object bipartite networks, which has found its applications in personalized recommendations. Huang *et al.* [25] analyzed the user-object networks (called *consumer-product networks* in Ref. [25]) to better understand the purchase behavior in e-commerce settings[1]. Grujić *et al.* [26, 27] studied the clustering patterns and degree correlations of user-movie bipartite networks according to the large-scale Internet Movie Database (IMDb), and applied a spectral analysis method to detect communities in the projected weighted networks. They found the monopartite networks for both users and movies exhibit an assortative behavior while the bipartite network shows a disassortative mixing pattern.

This Letter reports the empirical analysis on two well-known web sites, *audioscrobbler.com* and *del.icio.us*, where users are connected with music groups and bookmarks, respectively. Our main findings are threefold: (i) All the object-degree distributions are power-law, while the user-degree distributions obey stretched exponential functions. (ii) The networks exhibit disassortative mixing patterns, indicating that the fresh users tend to view popular objects and the unpopular objects are usually collected by very active users. (iii) We propose a new index, named *collaborative clustering coefficient*, to quantify the clustering behavior based on the collaborative connections. The two networks are of high collaborative clustering coefficients for both users and objects. For the lower-degree objects, a negative correlation between the object collaborative clustering coefficient and the object degree is observed, which disappears when the degree exceeds the average object degree. For audioscrobbler.com, the user collaborative clustering coefficient is strongly negatively correlated with the user degree, decaying in an exponential form for low degrees.

[1] Instead of the direct analysis on bipartite networks, Huang *et al.* [25] concentrated on the monopartite networks obtained from the bipartite networks.

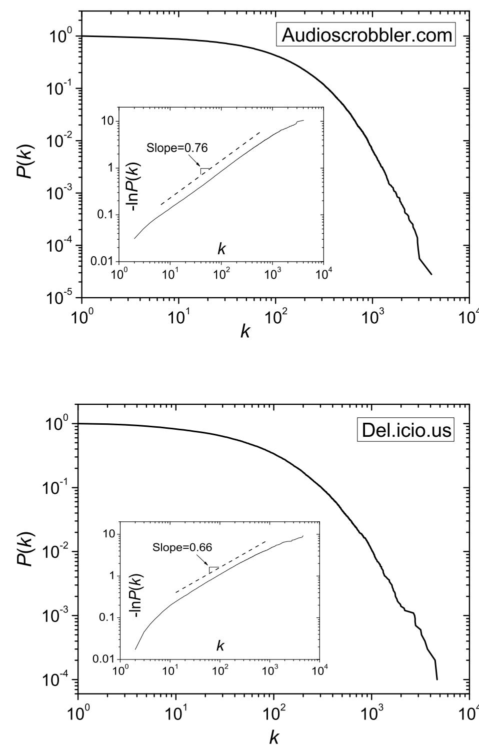

Fig. 2: Distributions of user degrees, which obey the stretched exponential form [31, 32]. We therefore plot the cumulative distribution $P(k)$ instead of $p(k)$ and show the linear fittings of $\log(-\log P(k))$ vs. $\log k$ in the insets.

Basic Concepts. – Figure 1 illustrates a small bipartite network that consists of six users and eight objects. The degree of user i, denoted by k_i, is defined as the number of objects connected to i. Analogously, the degree of object α, denoted by d_α, is the number of users connected to α. For example, as shown in Fig. 1, $k_i = d_\alpha = 3$. The density function, $p(k)$, is the probability that a randomly selected user is of degree k, while the cumulative function, $P(k)$, denotes the probability that a randomly selected user is of degree no less than k. The *nearest neighbors' degree* for user i, denoted by $d_{\mathtt{nn}}(i)$, is defined as the average degree over all the objects connected to i. For

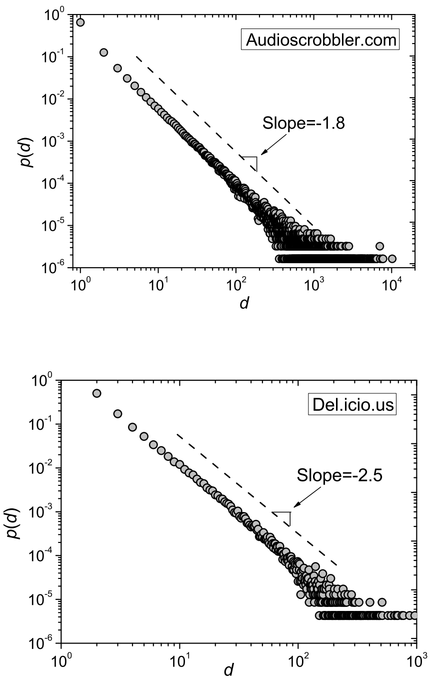

Fig. 3: Distributions of object degrees, which are power-law (they can pass the Kolmogorov-Smirnov test with threshold quantile 0.9) with exponents obtained by using the maximum likelihood estimation [33].

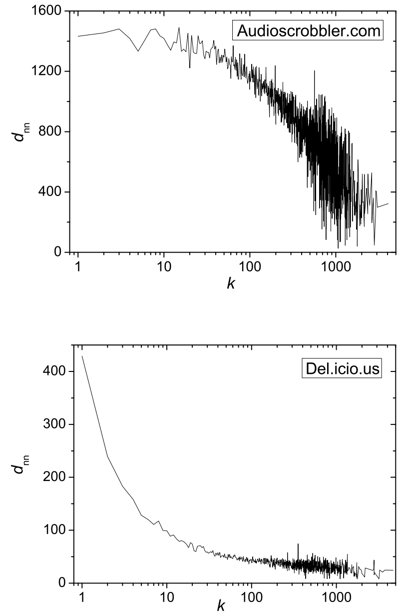

Fig. 4: The degree-dependent nearest neighbors' degree, $d_{\rm nn}(k)$, as a function of user-degree, k.

example, as shown in Fig. 1, $d_{\rm nn}(i) = \frac{d_\alpha + d_\beta + d_\gamma}{3} = \frac{7}{3}$. The degree-dependent nearest neighbors' degree, $d_{\rm nn}(k)$ is the average nearest neighbors' degree over all the users of degree k, that is, $d_{\rm nn}(k) = \langle d_{\rm nn}(i)\rangle_{k_i=k}$. Corresponding definitions for objects, say $p(d)$, $P(d)$, $k_{\rm nn}(\alpha)$ and $k_{\rm nn}(d)$, are similar and thus omitted here.

The traditional clustering coefficient [29] cannot be used to quantify the clustering pattern of a bipartite network since it always give a zero value. Lind *et al.* [11] proposed a variant counting the rectangular relations instead of triadic clustering, which can be applied to general bipartite networks. However, this Letter aims at a special class of bipartite networks, and thus we propose a new index to better characterize the clustering patterns resulted from the collaborative interests of users. A standard measure of object similarity according to the collaborative selection is the *Jaccard similarity* [30], $s_{\alpha\beta} = \frac{|\Gamma_\alpha \bigcap \Gamma_\beta|}{|\Gamma_\alpha \bigcup \Gamma_\beta|}$, where Γ_α and Γ_β are the sets of neighboring nodes of α and β, respectively. Obviously, $s_{\alpha\beta} = s_{\beta\alpha}$ and $0 \leq s_{\alpha\beta} \leq 1$ for any α and β. For example, as shown in Fig. 1, $s_{\alpha\beta} = s_{\beta\gamma} = \frac{1}{3}$ and $s_{\alpha\gamma} = \frac{1}{2}$. The *collaborative clustering coefficient* of user i is then defined as the average similarity between i's selected objects: $C_u(i) = \frac{1}{k_i(k_i-1)}\sum_{\alpha\neq\beta} s_{\alpha\beta}$, where α and β run over all i's neighboring objects. For example, as shown in Fig. 1, the collaborative clustering coefficient of user i is $C_u(i) = \frac{7}{18}$. The user collaborative clustering coefficient of the whole network is defined as $C_u = \frac{1}{N'}\sum_i C_u(i)$, where i runs over all users with degrees larger than 1 and N' denotes the number of these users. The degree-dependent collaborative clustering coefficient, $C_u(k)$, is defined as the average collaborative clustering coefficient over all the k-degree users. Corresponding definitions for objects are as following: (i) $C_o(\alpha) = \frac{1}{d_\alpha(d_\alpha-1)}\sum_{i\neq j} s_{ij}$, where $s_{ij} = \frac{|\Gamma_i \bigcap \Gamma_j|}{|\Gamma_i \bigcup \Gamma_j|}$ is the Jaccard similarity between users i and j; (ii) $C_o = \frac{1}{M'}\sum_\alpha C_o(\alpha)$, where M' denotes the number of objects with degrees larger than 1; (iii) $C_o(d)$ is the average collaborative clustering coefficient over all the d-degree objects.

Data. – This Letter analyzes two data sets. One is downloaded from audioscrobbler.com[2] in January 2005 by Lambiotte and Ausloos [23, 24], which consists of a listing of users, together with the list of music groups the users own in their libraries. Detailed information about this data set can be found in Refs. [23, 24]. The other is a random sampling of 10^4 users together with their collected bookmarks (URLs) from del.icio.us[3] in May 2008 [28]. Table 1 summarizes the basic statistics of these two data sets.

Empirical Results. – Figure 2 reports the degree distributions for users, which do not follow either the

[2]Audioscrobbler.com is a well-known collaborative filtering web site that allows user to create the personal web pages as their music libraries and to discover new music groups form other users' libraries.

[3]Del.icio.us is one of the most popular social bookmarking web sites, which allows users not only to store and organize personal bookmarks, but also to look into other users' collections and find what they might be interested in.

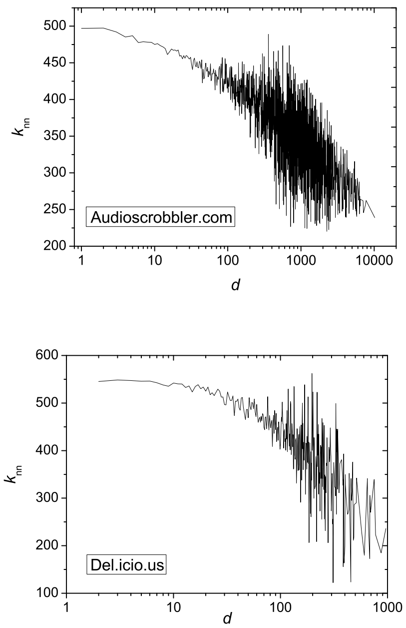

Fig. 5: The degree-dependent nearest neighbors' degree, $k_{\mathtt{nn}}(d)$, as a function of object-degree, d.

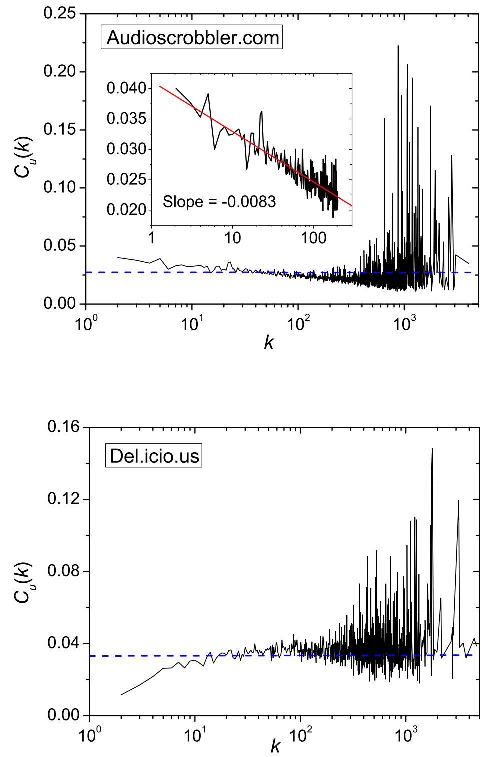

Fig. 6: (Color online) The clustering-degree correlations for users. Blue dash lines denote the collaborative clustering coefficients of the whole networks, C_u. The inset displays the early decaying behavior of $C_u(k)$ for audioscrobbler.com, which can be well fitted by an exponential form as $C_u(k) \sim \mathtt{e}^{-0.0083k}$.

power-law form or the exponential form. In fact, they lie in between exponential and power-law forms, and can be well fitted by the so-called *stretched exponential distributions* [31,32], as $p(k) \sim k^{\mu-1}\mathtt{exp}\left[-(\frac{k}{k_0})^{\mu}\right]$, where k_0 is a constant and $0 \leq \mu \leq 1$ is the characteristic exponent. The borderline $\mu = 1$ corresponds to the usual exponential distribution. For μ smaller than one, the distribution presents a clear curvature in a log-log plot. The exponent μ can be determined by considering the cumulative distribution $P(k) \sim \mathtt{exp}\left[-(\frac{k}{k_0})^{\mu}\right]$, which can be rewritten as $\mathtt{log}(-\mathtt{log}P(k)) \sim \mu\mathtt{log}k$. Therefore, Using $\mathtt{log}k$ as x-axis and $\mathtt{log}(-\mathtt{log}P(k))$ as y-axis, if the corresponding curve can be well fitted by a straight line, then the slope equals μ. Accordingly, as shown in Fig. 2, the exponents μ for audioscrobbler.com and del.icio.us are 0.76 and 0.66 respectively. These results have refined the previous statistics [23], where the exponential function is directly used to fit the user degree distribution of audioscrobbler.com. As shown in Fig. 3, all the object-degree distributions are power laws, as $p(d) \sim d^{-\phi}$. The exponents, ϕ, obtained by the maximum likelihood estimation [33], are shown in the corresponding figures.

As shown in Fig. 4 and Fig. 5, for both users and objects, the degree is negatively correlated with the average nearest neighbors' degree, exhibiting a disassortative mixing pattern. This result is in accordance with the user-movie bipartite network [26,27], indicating that the fresh users tend to view popular objects and the unpopular objects are usually collected by very active users. The correlation between $d_{\mathtt{nn}}$ and k is stronger than this between $k_{\mathtt{nn}}$ and d, which may be caused by the fact that the users are active while the objects are passive.

Table 1 reports the user collaborative clustering coefficients and object collaborative clustering coefficients for the whole networks. For comparison, we calculate the average user similarity over all user pairs, $\bar{s_u} = \frac{1}{N(N-1)}\sum_{i\neq j} s_{ij}$, and the average object similarity over all object pairs, $\bar{s_o} = \frac{1}{M(M-1)}\sum_{\alpha\neq\beta} s_{\alpha\beta}$. The connections for both users and objects are considered to be highly clustered since $C_u \gg \bar{s_o}$ and $C_o \gg \bar{s_u}$. The clustering-degree correlations for users are reported in Fig. 6. For audioscrobbler.com, a remarkable negative correlation for small-degree users is observed. Actually, $C_u(k)$ decays in an exponential form for small k. This result agrees with our daily experience that a heavy listener generally has broader interests of music[4]. In contrast, for del.icio.us a weakly positive correlation is observed for small-degree users. One reason for the difference between audioscrobbler.com and del.icio.us is that the collections in audio-

[4]In the statistical level, the collaborative clustering coefficient reflects the diversity of a user's tastes: the higher coefficient corresponds to the narrower tastes.

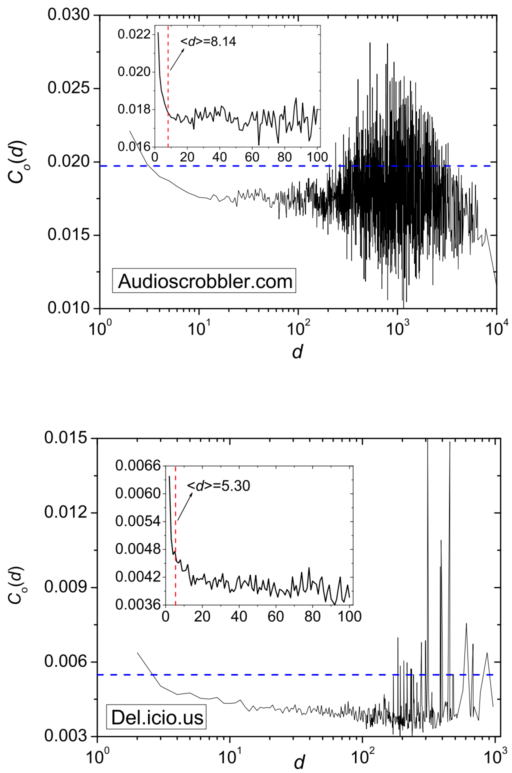

Fig. 7: (Color online) The clustering-degree correlations for objects. Blue dash lines denote the collaborative clustering coefficients of the whole networks, C_o. The insets display the early decaying behavior of $C_o(d)$, with the read dash lines denoting the average object degrees.

scrobbler.com only reflect the particular tastes of music, while the collections of URLs contain countless topics wherein music is just a very small one. In audioscrobbler.com, collections of a heavy listener (i.e., large-degree user) usually consist of several music genres, each of which contains a considerable number of music groups, while most of the music groups collected by a small-degree user belong to one genre. However, in del.icio.us, even for a very-small-degree user, his/her few collected URLs can be of highly diverse topics. Therefore, for del.icio.us, one can not infer that a small-degree user has limited interests. In addition, collections of music groups are mainly determined by personalized interests, while we have checked that in del.icio.us, many bookmarks are less personalized, that is, they can not well reflect the personal interests of users. For example, online tools like translators and search engines, and information services webs like the train schedules and air ticket centers are frequently collected. However, till now, we are not fully understood the origins of those nontrivial correlations, a future exploration making use of content-based or topic-based analysis on the URLs may provide a clearer picture.

Figure 7 reports the clustering-degree correlations for objects. For the lower-degree objects, a negative correlation between the object collaborative clustering coefficient and the object degree is observed, which disappears at about the average object degree. This result suggests that the unpopular objects (i.e., small-degree objects) may be more important than indicated by their degrees, since the collections of unpopular objects can be considered as a good indicator for the common interests–it is not very meaningful if two users both select a popular object, while if a very unpopular object is simultaneously selected by two users, there must be some common tastes shared by these two users. In fact, the empirical result clear shows that the users commonly collected some unpopular objects have much higher similarity to each other than the average. The information contained by those small-degree objects, usually having little effect in previous algorithms, may be utilized for better community detection and information recommendation.

Conclusion and Discussion. – Today, the exploding information confronts us with an information overload: we are facing too many alternatives to be able to find out what we really need. The collaborative filtering web sites provide a promising way to help us in automatically finding out the relevant objects by analyzing our past activities. In principle, all our past activities can be stored in the user-object networks (maybe in a weighted manner), which play the central role in those online services. This Letter reports the empirical analysis of two user-object networks based on the data downloaded from audioscrobbler.com and del.icio.us. We found that all the object-degree distributions are power-law while the user-degree distributions obey stretched exponential functions, which refines the previous results [23]. For both users and objects, the connections display disassortative mixing patterns, in accordance with the observations in user-movie networks [26, 27]. We proposed a new index, named collaborative clustering coefficient, to quantify the clustering behavior based on the collaborative selection. The connections for both users and objects are considered to be highly clustered since the collaborative clustering coefficients are much larger than the corresponding background similarities.

A problem closely related to the analysis of web-based user-object bipartite networks is how to recommend objects to users in a personalized manner [34, 35]. The empirical results reported in this Letter provide some insights in the design of recommendation algorithms. For example, as shown in Fig. 4, the average degree of collected objects is negatively correlated with the user's degree, and the fresh users tend to select very popular objects, that is, they have not well established their personalities and their collections are mostly popularity-based. This phenomenon gives an empirical explanation of the so-called *cold-start problem* [36], namely the personalized recommendations to the very-small-degree users are often inac-

curate. In addition, if we compare the significance of the user collaborative clustering coefficient, $C_u/\bar{s_o}$, and the significance of the object collaborative clustering coefficient, $C_o/\bar{s_u}$, we will find that for both audioscrobbler.com and del.icio.usm, the former (268.07 and 72.84) are much larger than the latter (4.11 and 6.79). Therefore, the fact that some users have commonly selected an object does not imply that they are much more similar to each other than two random users, however the objects selected by a user are statistically much more similar to each other than two random objects. The collaborative filtering techniques have two categories in general [34, 35]: one is user-based, which recommends to the target user the objects collected by the users sharing similar tastes; the other is object-based, which recommends the objects similar to the ones the target user preferred in the past. The comparison between $C_u/\bar{s_o}$ and $C_o/\bar{s_u}$ indicates that the object-based collaborative filtering will perform better, and such a kind of comparison can be considered as a helpful evidence before the choice between any user-based and object-based algorithms [37]. Furthermore, the clustering-degree correlations reported in Fig. 7 suggest that the small-degree objects actually play a more significant role than indicated by their degrees. In fact, we have already demonstrated that to emphasize the impacts of small-degree objects can remarkably enhance the recommendation algorithms' accuracies [38, 39]. We think the further in-depth analysis of information contained by the small-degree objects can find its applications in the design of more efficient and accurate recommendation algorithms.

* * *

We acknowledge the valuable suggestions and comments from Bosiljka Tadic and Renaud Lambiotte. This work was partially supported by Swiss National Science Foundation (grant no. 200020-121848), the National Natural Science Foundation of China under Grant Nos. 60973069 and 90924011. M.S.S. acknowledges the China Postdoctoral Science Foundation under Grant No. 20080431273 and the Sino-Swiss Science and Technology Cooperation (SSSTC) Project EG 20-032009. T.Z. acknowledges the National Natural Science Foundation of China under Grant Nos. 60744003 and 10635040.

REFERENCES

[1] R. Albert, A.-L. Barabási, Rev. Mod. Phys. 74 (2002) 47.
[2] S. N. Dorogovtsev, J. F. F. Mendes, Adv. Phys. 51 (2002) 1079.
[3] M. E. J. Newman, SIAM Rev. 45 (2003) 167.
[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Phys. Rep. 424 (2006) 175.
[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, P. R. U. Boas, Adv. Phys. 56 (2007) 167.
[6] P. Holme, F. Liljeros, C. R. Edling, B. J. Kim, Phys. Rev. E 68 (2003) 056107.
[7] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Åberg, Nature 411 (2001) 907.
[8] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabási, Nature 407 (2000) 651.
[9] P.-P. Zhang, K. Chen, Y. He, T. Zhou, B.-B. Su, Y.-D. Jin, H. Chang, Y.-P. Zhou, L.-C. Sun, B.-H. Wang, D.-R. He, Physica A 360 (2006) 599.
[10] Q. Xuan, F. Du, T.-J. Wu, Chaos 19 (2009) 023101.
[11] P. G. Lind, M. C. González, H. J. Herrmann, Phys. Rev. E 72 (2005) 056127.
[12] E. Estrada, J. A. Rodríguez-Velázquez, Phys. Rev. E 72 (2005) 046105.
[13] M. Peltomäki, M. Alava, J. Stat. Mech. (2006) P01010.
[14] R. Lambiotte, M. Ausloos, Phys. Rev. E 72 (2005) 066117.
[15] T. Zhou, J. Ren, M. Medo, Y.-C. Zhang, Phys. Rev. E 76 (2007) 046115.
[16] Y.-L. Wang, T. Zhou, J.-J. Shi, J. Wang, D.-R. He, Physica A 388 (2009) 2949.
[17] J. J. Ramasco, S. N. Dorogovtsev, R. Pastor-Satorras, Phys. Rev. E 70 (2004) 036106.
[18] J. Ohkubo, K. Tanaka, T. Horiguchi, Phys. Rev. E 72 (2005) 036120.
[19] M. L. Goldstein, S. A. Morris, G. G. Yen, Phys. Rev. E 71 (2005) 026108.
[20] J.-L. Guillaume, M. Latapy, Physica A 371 (2006) 795.
[21] J. Schafer, J. Konstan, J. Riedl, Data Min. &. Knowl. Discovery 5 (2001) 115.
[22] G. Bianconi, P. Laureti, Y.-K. Yu, Y.-C. Zhang, Physica A 332 (2004) 519.
[23] R. Lambiotte, M. Ausloos, Phys. Rev. E 72 (2005) 066107.
[24] R. Lambiotte, M. Ausloos, Eur. Phys. J. B 50 (2006) 183.
[25] Z. Huang, D. D. Zeng, H. Chen, Management Science 53 (2007) 1146.
[26] J. Grujić, Lect. Notes Comput. Sci. 5102 (2008) 576.
[27] J. Grujić, M. Mitrović, B. Tadić, *Proceedings of the 16th International Conference on Digital Signal Processing* (IEEE Press, 2009).
[28] Z.-K. Zhang, T. Zhou, Y.-C. Zhang, Physica A (2009) doi:10.1016/j.physa.2009.08.036.
[29] D. J. Watts, S. H. Strogatz, Nature 393 (1998) 440.
[30] P. Jaccard, Bulletin de la Societe Vaudoise des Science Naturelles 37 (1901) 547.
[31] J. Laherrère, D. Sornette, Eur. Phys. J. B 2 (1998) 525.
[32] T. Zhou, B.-H. Wang, Y.-D. Jin, D.-R. He, P.-P. Zhang, Y. He, B.-B. Su, K. Chen, Z.-Z. Zhang, J.-G. Liu, Int. J. Mod. Phys. C 18 (2007) 297.
[33] M. L. Goldstein, S. A. Morris, G. G. Yen, Eur. Phys. J. B 41 (2004) 255.
[34] J. L. Herlocker, J. A. Konstan, K. Terveen, J. T. Riedl, ACM Trans. Inf. Syst. 22 (2004) 5.
[35] G. Adomavicius, A. Tuzhilin, IEEE Trans. Knowl. Data Eng. 17 (2005) 734.
[36] A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock, *Proc. 25th Intl. ACM SIGIR Conf. Res. Develop. Inf. Retr.* (ACM Press, 2002).
[37] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, *Proc. 10th Intl. Conf. WWW* (ACM Press, 2001).
[38] T. Zhou, L. L. Jiang, R. Q. Su, Y.-C. Zhang, Europhys. Lett. 81 (2008) 58004.
[39] R.-R. Liu, C.-X. Jia, T. Zhou, D. Sun, B.-H Wang, Physica A 388 (2009) 462.